

Estimating the Mean Across Individuals of the Long-Term Proportion of a Nutrient Intake Contributed by a Food

Phillip S. Kott

Krebs-Smith, Guenther, and Kott (1989) show that the population mean of the proportion of a nutrient intake contributed by a food (say the fraction of cholesterol derived from eggs) can vary widely as the number of intake days under investigation increases. A stochastic model is proposed to explain this phenomenon. An individual's conceptual long-term nutrient-intake proportion for a specific food is defined here as the limit of his (her) D-day proportion as D grows arbitrarily large. A new estimator for this value based on a finite number of days of data is proposed. This naturally leads to an estimator for the population mean of individual long-term proportions. Unfortunately, a simulation reveals that the proposed estimator can be biased when based on the two-day intake data sets currently available. We conclude that the true population mean of individual long-term proportions is likely somewhere between the value of the proposed estimator and the commonly used "population proportion," in which the entire population of, say, N individuals is effectively treated as if it were a single individual with ND days of intake data, but that the proposed estimator tends to be less biased.

KEY WORDS: Population; bias; Correlation; Population proportion.

Phillip S. Kott is Chief Research Statistician, Research and Development Division, National Agricultural Statistics Service, 3251 Old Lee Highway, Room 305, Fairfax, Virginia, 22030. The author would like to thank Drs. Patricia Guenther and Susan Krebs-Smith for their support of this research and Lisa Kahle for did the actual computations. Comments on the methodology are welcomed by the author.

1. INTRODUCTION

Suppose one desires an estimate for the average across a population of individuals of the proportion of a nutrient's intake originating from a particular food. Krebs-Smith, Guenther, and Kott (1989) have shown empirically that estimated "mean proportions" (their terminology) can vary widely as the number of intake days increases. This note mathematically explains that phenomenon. It also proposes a new estimator for the mean proportion which is renamed, less ambiguously, the *mean individual long-term proportion*.

Section 2 lays out the conceptual framework. Section 3 describes the new estimator for the mean individual long-term proportion and contrasts it with the commonly-estimated population proportion, which effectively treats the entire population as a single individual. Section 4 computes the proposed estimator for the mean individual long-term proportion and the estimated population proportion for five nutrient/food combinations. A simulation suggests that neither is a completely satisfying estimator for the true mean individual long-term proportion. Section 5 contains a discussion.

2. AN INDIVIDUAL LONG-TERM PROPORTION

Although our interest is in estimating a population characteristic, it is helpful to begin by focusing attention on a single individual. We first need to define the long-term nutrient-intake proportion of a specific food for this individual assuming a conceptual framework under which the underlying eating behavior of the individual does not change over time. From there we can define the mean individual long-term proportion for a population containing that individual. Following Krebs-Smith, Guenther, and Kott (1989), we will use the guiding example of the proportion of cholesterol derived from eggs. It is a simple matter to apply the analysis to other nutrient-intake proportions for specific foods.

Let T_{kd} denote the cholesterol intake of individual k on day d , Y_{kd} the individual's

amount of cholesterol derived from eggs on that day, and $R_{kd} = Y_{kd}/T_{kd}$ the proportion of cholesterol derived from eggs by k on d . We will assume that T_{kd} and R_{kd} obey the following stochastic model:

$$\begin{aligned} T_{kd} &= m_k(1 + \tau_{kd}) \\ R_{kd} &= \mu_k(1 + \epsilon_{kd}), \end{aligned} \tag{1}$$

where τ_{kd} and ϵ_{kd} are, respectively, identically distributed random variables with mean zero and finite higher moments. In addition, we assume both random variables are serially independent; that is, independent across days and across individuals. On the other hand, we let $E(\epsilon_{kd}\tau_{kd}) = \text{Cov}_k(\epsilon, \tau)$; that is, we allow that possibility that ϵ_{kd} and τ_{kd} are correlated for the same day and individual.

A more complete model of individual intake behavior would allow T_{kd} and R_{kd} to be functions of a variety of characteristics (the day of the week, the season, etc.). The simple model in equation (1), however, serves our present purpose well.

Suppose we have intakes for the individual k for D non-consecutive days. Let

$$\begin{aligned} R_{k(D)} &= \frac{\sum_{d=1}^D Y_{kd}}{\sum_{d=1}^D T_{kd}} \\ &= \frac{\sum_{d=1}^D T_{kd} R_{kd}}{\sum_{d=1}^D T_{kd}} \end{aligned}$$

denote the proportion of the D -day intake of cholesterol that was derived from eggs.

The individual k long-term proportion of cholesterol from eggs can be defined as the probability limit of $R_{k(D)}$ as D grows arbitrarily large:

$$R_{k(\infty)} = \lim_{D \rightarrow \infty} R_{k(D)} = \mu_k[1 + \text{Cov}_k(\epsilon, \tau)], \quad (2)$$

where the second equality is demonstrated in the appendix. The appendix also shows that the expectation of $R_{k(D)}$ for a particular finite D is approximately

$$E[R_{k(D)}] \approx \mu_k\{[1 + \text{Cov}_k(\epsilon, \tau)][1 - 1/k]\} \quad (3)$$

under strong assumptions we *provisionally* assume to hold.

An individual's daily proportion of cholesterol from eggs is correlated with his (her) daily intake of cholesterol (Krebs-Smith, Guenther, and Kott 1989). This means, $\text{Cov}_k(\epsilon, \tau) > 0$. When $\text{Cov}_k(\epsilon, \tau) \neq 0$, equation (3) reveals that $R_{k(D)}$ is a biased estimator of $R_{k(\infty)}$ for any finite D . This bias gets smaller in absolute terms as D gets larger. When D is very large (greater than 10 should work comfortably in practice), the distinction between $R_{k(D)}$ and $R_{k(\infty)}$ is very small.

When D is small but greater than 1, a nearly unbiased estimator of the individual k 's long-term proportion, $R_{k(\infty)}$, under our strong provisional assumptions is easily established; namely,

$$r_{k(\infty); D} = (D/[D - 1])R_{k(D)} - (1/[D - 1]) \sum_{d=1}^D R_{kd}/D. \quad (4)$$

One should be aware that although $r_{(\infty, k)}$ is a nearly unbiased estimator of the long-term proportion for k under certain assumptions, it may not be a very good estimator. In fact, it may even be negative. This is of limited concern here, however, because our real goal is to estimate the mean of the individual long-term nutrient-intake proportions for specific foods across a population of individuals.

3. ESTIMATING THE MEAN INDIVIDUAL LONG-TERM PROPORTION

We can define the *mean individual long-term nutrient-intake proportion* for specific food within a population of N individuals as

$$\begin{aligned} R_{(\infty)} &= \sum_{k=1}^N R_{k(\infty)} / N \\ &= \sum \mu_k [1 + \text{Cov}_k(\epsilon, \tau)] / N \end{aligned} \quad (5)$$

Suppose we have intake information for D non-consecutive days from a random sample of n individuals in the population. A nearly unbiased estimator for $R_{(\infty)}$ based on that information under our strong provisional assumptions is

$$r_{(\infty; D)} = \sum_{k=1}^n w_k r_{k(\infty, D)} / \sum_{k=1}^n w_k, \quad (6)$$

where w_k is the sampling weight for individual k (which is proportional to the inverse of his (her) selection probability, perhaps adjusted for nonresponse).

An alternative estimator of the mean individual long-term proportion is the estimated *population proportion* based on D days of intake data:

$$\begin{aligned} p_{(D)} &= \sum_{k=1}^N w_k \sum_{d=1}^D Y_{kd} / \sum_{k=1}^N w_k \sum_{d=1}^D T_{kd} \\ &= \sum w_k \sum T_{kd} R_{kd} / \sum w_k \sum T_{kd} \end{aligned}$$

It is not hard to see that $p_{(D)}$ is a good estimator for the mean individual long term proportion when m_k , μ_k and $\text{Cov}_k(\epsilon, \tau)$ are each constant across all individuals in the

population, and n is large (say larger than $10/D$). This is because the constancy of the three parameters would allow us to treat the population as a single individual with ND days of nutrient intake.

Now $p_{(D)}$ is a consistent estimator for

$$P_{(D)} = \frac{\sum_{k=1}^N \sum_{d=1}^D T_{kd} R_{kd}}{\sum_{k=1}^N \sum_{d=1}^D T_{kd}} \quad (7)$$

under mild conditions. The right-hand side of equation (7) is approximately equal to $\sum^N m_k \mu_k [1 + \text{Cov}_k(\epsilon, \tau)] / \sum^N m_k$ when N is large. This in turn equals $\sum^N \mu_k [1 + \text{Cov}_k(\epsilon, \tau)] / N = R_{(\infty)}$ when all the m_k are the same. Thus, $p_{(D)}$ would be a good estimator for the mean individual long-term proportion, $R_{(\infty)}$, if the long-term average daily intake of cholesterol, m_k , alone were constant across all individuals in the population.

Unfortunately, in the real world, the long-term average daily intake of cholesterol varies across individuals in a population. When those with higher than average daily intakes also tend to derive a higher proportion of their daily cholesterol from eggs (μ_k for individual k), then $p_{(D)}$ will tend to overestimate $R_{(\infty)}$.

4. AN EMPIRICAL INVESTIGATION

Using two-nonconsecutive-day-intake data from the 1994-96 Continuing Survey of Intakes by Individual (CSFII), we estimated the following proportions among adults 20 years or older:

Cholesterol from eggs

Vitamin A from dark green or deep yellow vegetables

Vitamin C from citrus fruits and juices

Energy from soft drinks

Calcium from milk and milk products.

We restricted our attention to the 9,159 adults who had positive values for calories, calcium, cholesterol, vitamin A and vitamin C over the two days. We used the two-day survey weights in computing the results displayed in Table 1. Similar results were found for men and women when analyzed separately and for children.

All estimates were computed using both days of intake data to control the potential impact of sequence effects (i.e., systematic differences in what is reported by an individual on the first and second days of a survey). That being said, we will ignore the possibility of sequence or any other type of measurement bias for the remainder of this analysis.

Four of the five nutrient/food combinations exhibit the same pattern. The estimated mean one-day proportion, $r_{(1; 2)}$, is lowest, followed by, $r_{(2)}$, the estimated mean two-day proportion, then, $r_{(\infty; 2)}$, the estimated mean long-term proportion, and finally, $p_{(2)}$, the estimated population proportion. The one exception is energy from soft drinks, for which only the population proportion is noticeably higher than the others,

roughly 5.0% as opposed to 4.7 or 4.8%. For cholesterol from eggs, by contrast, the estimates range from $r_{(1; 2)}$ at 15.7% to $p_{(2)}$ at 28.2%, increasing at roughly constant intervals through $r_{(2)}$ (19.6%) and $r_{(\infty; 2)}$ (23.4%).

If $r_{(\infty; 2)}$ is an unbiased estimator for the mean long-term nutrient-intake proportion for a specific food, then it would appear that $p_{(2)}$ is biased upward (i.e., tends to be too large) as an estimator for $R_{(\infty)}$ for all five combinations. For cholesterol from eggs, for example, the population proportion may be as biased ($28.2 - 23.4 = 4.8\%$) as using the two-day proportion directly ($19.6 - 23.4 = -4.8\%$), only in the opposite direction.

Unfortunately, $r_{(\infty; 2)}$ may itself be biased as an estimator for $R_{(\infty)}$. To assess the impact of possible bias, we treat all the survey data as if they came from a single individual. If this were the case, then the unweighted estimated population proportion based on $9,159 \times 2$ days of intake data would be an unbiased estimator for the true mean long-term proportion.

We computed the four estimators ignoring the weights, first, with each person's intake as (s)he provided them, and then with the first-intake day of each individual matched with the second-intake day from a randomly selected individual. Selections were made without replacement.

Table 2 displays the results. The first line associated with a nutrient/food combination was computed with the originally-matched data and the second line with the randomly-matched data. The first produces estimates as if the individuals were differentiated. The second as if they were not; that is, as if all the data came from a single person.

Observe that the mean estimated one-day proportions are the same using either method and are the estimated population proportions. This comes as no surprise given their natures and the laws of arithmetic.

When the data is treated as if it comes from a single individual, $r_{(\infty; 2)}$ appears to

be downward biased as an estimator for $R_{(\infty)}$ for all five combinations; that is to say, it always smaller than the unbiased $p_{(2)}$. This bias of $r_{(\infty; 2)}$ is negligible for energy from soft drinks (both $r_{(\infty; 2)}$ and $p_{(2)}$ are roughly 4.6%) and calcium from milk (47%). For vitamin C from citrus, the bias is relatively small. It is more pronounced for cholesterol from eggs and vitamin A from vegetables.

The estimate $r_{(\infty; 2)}$ (like $r_{(2)}$) is always less when computed with the originally-matched data set rather than the randomly-matched set. This suggests that one cannot assume that every individual in the population really has the same underlying behavior. As a consequence, the population proportion appears to be biased upward as an estimator for the mean long-term proportion, at least, for energy from soft drinks and calcium from milk and very likely for all five combinations.

5. DISCUSSION

When one has only two days of intake data per individual, and the estimated mean one-day proportion of the contribution of particular food to a certain nutrient is roughly the same, as the estimated mean two-day proportion, it scarcely matters how the mean long-term proportion is estimated – as long as the estimated population proportion is not used for that purpose. When the two estimates do vary meaningfully, as they do in four of the five nutrient/food combinations analyzed in the previous section, and the estimated mean two-day proportion exceeds the estimated mean one-day proportion, then both are likely biased downward. This is because an individual's daily proportion of the nutrient contributed by the food is correlated his (her) daily intake of the nutrient.

In the text we developed a new estimator for the mean long-term nutrient-intake proportion for a specific food based on two days of intake data. Sadly, it too proved to

be often biased downward, albeit by a trivial amount for some nutrient/food combinations. We feel it is almost always at least as good an estimator for the mean long-term proportion as the estimated population proportion, which has a tendency to be biased upward. When the estimated population proportion for a nutrient/food combination exceeds the estimated mean one-day proportion by a meaningful amount, those individuals in the population with higher-than-average daily intakes of the nutrient tend to have higher-than-average daily nutrient-intake proportions derived from that food causing the upward bias in the population proportion.

Unfortunately, we could not come up with a useful indicator for how much the new estimator would be biased downward, except when the estimated mean one and two-day intake proportions are roughly equal, and the biases from all three are trivial.

APPENDIX: Proofs of equations (2) and (3)

From its definition,

$$\begin{aligned}
 R_{k(D)} &= \frac{\sum_{d=1}^D T_k R_k / \sum_{d=1}^D T_k}{\sum_{d=1}^D m_k(1 + \tau_{kd})\mu_k(1 + e_{kd})/D} \\
 &= \mu_k \frac{1 + \sum_{d=1}^D \tau_{kd}/D + \sum_{d=1}^D e_{kd}/D + \sum_{d=1}^D e_{kd}\tau_{kd}/D}{1 + \sum_{d=1}^D \tau_{kd}/D}.
 \end{aligned}$$

After some manipulation,

$$\begin{aligned}
 R_{(D)} &= \mu_k \left(1 + \sum_{d=1}^D e_{kd}/D - \sum_{d=1}^D e_{kd}\tau_{kd}/D - [\sum_{d=1}^D e_{kd}/D][\sum_{d=1}^D \tau_{kd}/D] + \right. \\
 &\quad \left. \mu_k \frac{(\sum_{d=1}^D \tau_{kd}/D)^2(\sum_{d=1}^D e_{kd}/D) - (\sum_{d=1}^D \tau_{kd}/D)(\sum_{d=1}^D e_{kd}\tau_{kd}/D)}{1 + \sum_{d=1}^D \tau_{kd}/D} \right). \tag{A1}
 \end{aligned}$$

Since e_{kd} and τ_{kd} have finite higher moments and are serially independent, the probability limit of both $\sum^D \tau_{kd}/D$ and $\sum^D e_{kd}/D$ is 0 as D grows arbitrarily large, while the plim of $\sum^D e_{kd}\tau_{kd}/D$ is $\text{Cov}_k(e, \tau)$. As a result $R_{(\infty)} = \text{plim}_{D \rightarrow \infty} R_{(D)} = \mu_k[1 + \text{Cov}_k(e, \tau)]$ as stated in equation (2).

As for the expectation of $R_{k(D)}$ for a given D , note first that the last line in equation (A1) exactly equals 0 when $D = 1$. When D exceeds 1, we need to assume this expression is close enough to zero in expectation to be ignored for equation (3) to

follow when expectations are taken on both sides of equation (A1). This strong assumption is reasonable when D is sufficiently large or when ϵ_i and τ_i have sufficiently small higher-order moments. Unfortunately, D is not small in practice. It is most often only two. Moreover, the higher order moments of ϵ_{kd} and τ_{kd} can be uncomfortably large in practice, as we have indirectly seen in the text.

REFERENCES

KREBS-SMITH S., GUENTHER P., AND KOTT, P. (1989). Mean proportion and population proportion: two answers to the same question? *J. Am. Diet. Assoc.* **89**, 671-6.

U.S. DEPARTMENT OF AGRICULTURE (1987). Nationwide Food Consumption Survey, Continuing Survey of Food Intakes By Individuals, Women 19-50 Years and their Children 1-5 Years, 4 Days, 1985. Hyattsville, MD: Human Nutrition Information Service, USDA.

U.S. DEPARTMENT OF AGRICULTURE (1988). Nationwide Food Consumption Survey, Continuing Survey of Food Intakes By Individuals, Women 19-50 Years and their Children 1-5 Years, 4 Days, 1986. Hyattsville, MD: Human Nutrition Information Service, USDA.

Table 1. Estimates for the Mean Individual Long-Term Proportions Among Adults 20 and Over

<i>Nutrient</i>	<i>Food</i>	<i>Estimators¹</i>			
		$r_{(1; 2)}$	$r_{(2)}$	$r_{(\infty; 2)}$	$p_{(2)}$
Cholesterol	Eggs	0.15728	0.19552	0.23376	0.28185
Vitamin A	Dark green and deep yellow vegetables	0.18477	0.23109	0.27740	0.32943
Vitamin C	Citrus fruits and juices	0.16270	0.18861	0.21452	0.28089
Energy	Soft drinks	0.04753	0.04710	0.04668	0.04977
Calcium	Milk	0.39183	0.41821	0.44459	0.48297

¹ *Notation for the estimators*

$r_{(1; 2)} = \sum w_k(R_{k1} + R_{k2})/2$ Estimated mean one-day proportion from two days of intake data

$r_{(2)} = \sum [w_k(T_{k1}R_{k1} + T_{k2}R_{k2})/(T_{k1} + T_{k2})]$ Estimated mean two-day proportion

$r_{(\infty; 2)} = 2r_{(2)} - r_{(1)}$ Estimated mean long-term proportion from two days of intake data

$p_{(2)} = \sum w_k(T_{k1}R_{k1} + T_{k2}R_{k2}) / \sum w_k(T_{k1} + T_{k2})$ Estimated population proportion from two days of intake data

Table 2. Unweighted Estimates for the Mean Individual Long-Term Proportions Among Adults 20 and Over and the Same Estimates with Random Matches of Day-One and Day-Two Data

<i>Nutrient</i>	<i>Food</i>	<i>Estimators</i> ¹			
		$r_{(1; 2)}$	$r_{(2)}$	$r_{(\infty; 2)}$	$p_{(2)}$
Cholesterol	Eggs	0.16722	0.20757	0.24793	0.30038
		0.16722	0.21842	0.26962	0.30038
Vitamin A	Dark green and deep yellow vegetables	0.18496	0.23022	0.27549	0.32929
		0.18496	0.23818	0.29141	0.32929
Vitamin C	Citrus fruits and juices	0.15940	0.18321	0.20702	0.27208
		0.15940	0.20956	0.25972	0.27208
Energy	Soft drinks	0.04339	0.04303	0.04267	0.04575
		0.04339	0.04453	0.04566	0.04575
Calcium	Milk	0.38188	0.40721	0.43254	0.47243
		0.38188	0.42616	0.47045	0.47243

¹ Notation is the same as in Table 1 with all w_k set equal to 1. The first line matches the day-one and day-two for the same individual. The second line uses random matches.